CerraData-4 MM: A Multimodal Benchmark Dataset on Cerrado for Land Use and Land Cover Classification

Mateus de Souza Miranda, Ronny Hänsch, Valdivino Alexandre de Santiago Júnior, Thales Sehn Körting, and Erison Carlos dos Santos Monteiro

***Abstract*—The *Cerrado* faces increasing environmental pressures, necessitating accurate land use and land cover mapping despite challenges, such as class imbalance and visually similar categories. To address this, we present CerraData-4 MM, a multimodal dataset combining Sentinel-1 synthetic aperture radar and Sentinel-2 multispectral imagery with 10 m spatial resolution. The dataset includes two hierarchical classification levels with seven and 14 classes, respectively, focusing on the diverse *Bico do Papagaio* ecoregion. We benchmark two models trained on CerraData-4 MM, employing a visual transformer-based architecture and a convolutional-based architecture. The ViT achieves superior performance in multimodal scenarios, with the highest macro F1-score of 57.60% and a mean Intersection over Union of 49.05% at the first hierarchical level. Both models struggle with minority classes, particularly at the second hierarchical level, where U-Net's performance drops to an F1-score of 18.16%. Weighted loss improves representation for underrepresented classes but reduces overall accuracy, underscoring the trade-off in weighted training. CerraData-4 MM offers a challenging benchmark for advancing deep learning models to handle class imbalance and multimodal data fusion.**

***Index Terms*—Cerrado, deep learning (DL), hierarchical level of classes, land use and land cover (LULC) classification, semantic segmentation.**

I. Introduction

The Brazilian *Cerrado* is renowned as the most biodiverse savanna in the world [12], featuring vegetation highly adapted to prolonged dry seasons. Spanning a significant portion of Brazil's territory, the *Cerrado* is divided into ecoregions, each defined by distinct characteristics shaped by their geographical location and geological formations, contributing to its diverse vegetation [13]. In addition, this biome is critical for agricultural production, accounting for over 55% of the Brazil's food production [14] and is the source of 8 out of 12 principal hydrological basins of Brazil [15]. Despite its ecological and economic importance, accurate land use and land cover (LULC) mapping in the Cerrado remains challenging due to the biome's heterogeneity and dynamic land-cover changes driven by human activities.

The TerraClass program [16], [17] plays an essential role in mapping LULC changes in the *Cerrado* region and identifies areas that have undergone deforestation and subsequent conversion to agriculture, pastures, secondary vegetation, and other types of land cover. According to the program, agricultural areas have expanded, primarily at the expense of pasture. However, the establishment of new pasture areas remains the main driver of natural vegetation conversion in the *Cerrado* biome. These maps support detailed analyses of changes in vegetation cover every two years. The TerraClass LULC map is created using a semi-automatic LULC classification and relies on manual visual interpretation and annotation as the final refinement, which is time-consuming and expensive. On the other hand, deep learning (DL) approaches are promising alternatives in terms of efficiency and agility, such as semantic segmentation models, e.g., vision transformer-based architectures [18] and convolutional-based models [19], to assist in the LULC classification.

Current LULC datasets for *Cerrado* are either geographically restricted or oversimplified in their class representations. For instance, some focus on small protected areas with few classes [2], while others, although detailed, are limited to specific localities, such as Brasília's National Park approached in [20]. These datasets are often exclusive to individual studies, preventing broader reproducibility. Moreover, most rely solely on optical data, which is susceptible to weather conditions and does not capture structural and moisture-related features critical to distinguishing *Cerrado* vegetation and agricultural types. synthetic aperture radar (SAR) data, with its all-weather capability and sensitivity to surface structure and moisture, offer a complementary solution [21]. However, no publicly available dataset integrates SAR and multispectral imagery (MSI) for the *Cerrado*, leaving a critical gap in multimodal LULC research.

To address these challenges, this paper presents CerraData-4 MM, the first multimodal dataset designed for the Bico do Papagaio ecoregion of the *Cerrado*, combining SAR and

Received 5 February 2025; revised 12 May 2025; accepted 28 June 2025. Date of publication 3 July 2025; date of current version 29 July 2025. This work was supported in part by theSão Paulo Research Foundation (FAPESP) under Grant 2023/09118-6, in part by the Brazilian National Council for Scientific and Technological Development (CNPq) under Grant 302205/2023-3 and Grant 140459/2023-5, and in part by the Brazilian Federal Agency for Support and Evaluation of Graduate Education through the Institutional Program for Internationalization (PrInt-CAPES) under Grant 88887.936146/2024-00.
(Corresponding author: Valdivino Alexandre de Santiago Júnior.)

Mateus de Souza Miranda, Valdivino Alexandre de Santiago Júnior, Thales Sehn Körting, and Erison Carlos dos Santos Monteiro are with Laboratório de Inteligência ARtificial para Aplicações AeroEspaciais e Ambientais (LIAREA), Programa de Pós-Graduação em Computação Aplicada (PGCAP), Instituto Nacional de Pesquisas Espaciais (INPE), São José dos Campos 12227-010, Brazil (e-mail: valdivino.santiago@inpe.br).

Ronny Hänsch is with German Aerospace Center (DLR), Oberpfaffenhofen 82234, Germany.

Digital Object Identifier 10.1109/JSTARS.2025.3585805

TABLE I
OVERVIEW OF MULTIMODAL REMOTE SENSING DATASETS AND THEIR KEY CHARACTERISTICS

Dataset	**Nr. of images**	**Image size (px)**	**Spatial resolution (m)**	**Time series**	**SAR**	**Optical**	**Others**	**Nr. of classes**	**Region**
SEN12MS [1]	541,986	256×256	10	No	Yes	MSI	-	33	Globally distributed
Brazilian *Cerrado*-Savanna Scenes [2]	1,311	64×64	5	No	No	NIRGB	-	4	*Serra do Cipó region*, on *Cerrado* biome
MultiSenGE [3]	8,157	256×256	10	Yes	Yes	MSI	-	14	France
MDAS [4]	6268	1,371×888, 1,371×888, 1,170×765	0.25, 10 and 30	No	Yes	HSI, MSI	DSM	14	Augsburg, Germany
DFC20 [5]	180,662	256×256	10	No	Yes	MSI	-	10	Globally distributed
GeoNRW [6]	7,782	1,000×1,000	1	No	Yes	MSI	DEM	10	North-Rhine, Westphalia
PASTIS-R [7]	2433 times series	128×128	10	Yes	Yes	MSI	-	18	French metropolitan
DFC22 [8]	5,016	2,000×2,000	0.5	No	No	RGB	DEM	L1: 12; L2: 14	France
FLAIR-one [9]	77,412	512×512	0.2	Yes	No	MSI	DSM, DTM	13	French metropolitan
Hunan Multimodal Dataset [10]	450	256×256	10	No	Yes	MSI	SRTM	7	Hunan, China
C2Seg [11]	4 scenes	2465×811, 886×1360, 13474×8706, 6225×8670	10	No	Yes	HSI, MSI	-	13	Berlin-Augsburg (Germany) and Beijing–Wuhan (China)
CerraData-4MM	30,322	128×128	10	No	Yes	MSI	-	Level-1: 7, level-2: 14	Bico do Papagaio ecoregion

MSI data with meticulously annotated reference masks. This dataset introduces a two-level hierarchical classification scheme: the first level comprises seven macro-classes, e.g., agriculture, pasture, natural vegetation, whereas the second level refines these into 14 subclasses, including specific crop types, stages of forest regeneration, and urban features. This hierarchical level of classes enables studies on fine-scale land-cover dynamics, such as pasture-to-crop transitions and secondary vegetation recovery, key processes in the *Cerrado*'s ongoing transformation. By openly providing this resource, this paper aims to 1) facilitate the development of robust DL models for LULC mapping in savanna ecosystems, 2) promote reproducibility in remote sensing research, and 3) demonstrate the synergistic potential of SAR and MSI data for capturing the *Cerrado*'s unique land-cover patterns.

II. RELATED DATASETS

Establishing and openly sharing multimodal datasets in the remote sensing field is crucial for advancing both technological development and practical applications. This section introduces several representative datasets designed for LULC classification using multimodal data, particularly SAR and MSI. Table I highlights some of the most state-of-the-art resources in this domain. These datasets differ in scope, spatial resolution, modalities, geographic focus, and classification tasks, addressing various research objectives and applications. Some works offer a rich diversity of samples, as data are collected from different parts of the world, such as SEN12MS and DFC2022, making them suitable for general-purpose studies. In contrast, others are designed for regional and specific areas of interest, providing high-resolution data for localized analyses, such as GeoNRW and the Hunan Multimodal Dataset.

The SEN12MS dataset [1] features globally distributed samples, offering a high diversity of characteristics in 33 classes, with a strong emphasis on natural categories such as various types of forests and savanna formations in all seasons. The dataset has 180 662 patch triplets, including dual-polarization SAR, MSI, and MODIS land-cover maps. The spatial resolution of the LULC maps (500 m), as well as the satellite imagery (60, 20, and 10 m), is resampled to a resolution of 10 m. The reference maps consequently exhibit low quality in terms of the richness of detail within the polygons.

MultiSenGE [3], in contrast, focuses on the eastern region of France and integrates SAR and MSI time series. Although it contains 8157 patches, each measuring 256× 256 pixels, collected from the eastern region of France. In addition, it includes time series data for both modalities. Due to the limited geographical coverage, the dataset's classes predominantly represent urban and agricultural landscapes, such as vineyards and orchards. This specificity renders the dataset highly tailored to a particular application context, with constrained sample diversity and geographic scope.

The C2Seg [11] dataset is designed to evaluate semantic segmentation performance across cities. It integrates three remote sensing modalities: hyperspectral, multispectral, and SAR, across two cross-city pairs: Berlin–Augsburg (Germany) and Beijing–Wuhan (China). C2Seg's strength lies in its modal diversity and domain generalization challenge, making it ideal for benchmarking multimodal learning and adversarial domain adaptation. The dataset includes semi-automatically annotated land cover classes from OpenStreetMap, covering 13 categories, all resampled to a spatial resolution of 10 m.

In particular, MultiSenGE, PASTIS-R, and FLAIR-one incorporate time series data, enabling temporal analysis. In contrast, SEN12MS and DFC20 consist solely of single snapshot images. Furthermore, MDAS and DFC22 utilize complementary data, such as digital surface models (DSM) and digital terrain models (DTM), alongside SAR and MSI images. Despite the availability of various data modalities that support model development for various remote sensing applications, many datasets focus primarily on land use classes, such as buildings, airports, and roads. SEN12MS is a notable exception, offering a comprehensive list of 33 classes related to vegetation types, including urban areas and vegetation domains.

When it comes to *Cerrado*, the Brazilian *Cerrado*-Savanna Scenes dataset [2] is an early attempt to provide labeled imagery. This dataset includes only 1311 samples across four LULC classes in the *Serra do Cipó* region, covering three main groups of types of natural land cover and agricultural areas. The dataset exhibits a limited representation of class samples due to its restricted diversity. In addition, the images are limited to capturing only the green, red, and near-infrared bands. In contrast, CerraData-4 MM offers more spectral bands, additional data modalities (SAR), and a much larger spatial coverage, offering 30 322 patches.

Every dataset fulfill distinct but synergistic roles in remote sensing research. However, a critical limitation shared by many existing datasets is the quality and accuracy of reference labels. For instance, SEN12MS [1] employs resampled MODIS land cover maps at 500 m resolution, resulting in imprecise polygon boundaries that reduce their utility for fine-grained analysis. Furthermore, most datasets prioritize generalized land use categories, e.g., "urban" or "cropland," over more fine-grained natural land-cover distinctions, limiting their applicability for biome-specific studies requiring transferability across regions. CerraData-4 MM, on the other hand, excels in label accuracy, addressing these shortcomings by providing manually annotated reference masks at 10 m spatial resolution, with a class hierarchy specifically designed to capture the *Cerrado*'s heterogeneous landscapes, from agricultural subtypes to regeneration stages, while maintaining rigorous labeling process.

CerraData-4 MM represents a unique contribution as the first publicly available dataset to integrate Sentinel-1 (SAR) and Sentinel-2 (MSI) imagery specifically for the *Cerrado* biome. This multimodal approach contributes to LULC monitoring, enabling researchers to overcome persistent cloud cover and seasonal variability that hinder optical-only solutions. While benchmark datasets like SEN12MS offer broader geographical coverage, CerraData-4 MM's regional focus provides unmatched granularity for studying the *Cerrado*'s rapid land-use transitions, where agricultural expansion, e.g., soybean and cotton cultivation, increasingly displaces natural vegetation. The dataset's value lies not in superseding global resources, but in enabling targeted studies of this ecologically vulnerable savanna.

The CerraData series [22] has progressively advanced LULC research for the *Cerrado*, with earlier versions establishing foundational resources: CerraData-1 delivered 1.5 million unlabeled CBERS-4 A images (2 m resolution) for self-supervised learning, while CerraData-2/3 introduce 130 000 labeled samples for scene classification [23]. CerraData-4 MM extends this lineage with three key innovations: 1) Sentinel-1 (S1) and 2 (S2) integration for multimodal analysis, 2) a hierarchical labeling system, comprising seven macro-classes and 14 subclasses, supporting both general and fine-grained studies, and 3) focus on the Bico do Papagaio ecoregion, a deforestation hotspot where Amazon-*Cerrado* transitions amplify classification challenges. With 30 322 patches of 128 × 128 px at 10 m resolution, the dataset captures spectral–structural complexities unique to this region, particularly in distinguishing crop types and secondary vegetation stages that exhibit overlapping patterns.

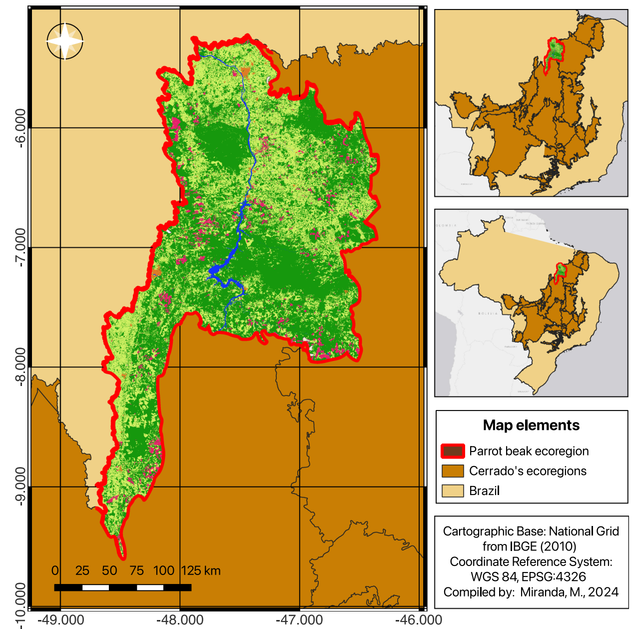

Fig. 1. *Bico do Papagaio*, marked in brown with red edges, is the region selected to create the dataset.

III. Dataset

CerraData-4 MM is a multimodal dataset crafted specifically for the *Cerrado* biome, focusing on the *Bico do Papagaio* ecoregion—one of the 19 officially recognized subdivisions of the biome [13]. As illustrated in Fig. 1, this region is situated in the northernmost area of the *Tocantins* state and shares a transitional boundary with the Amazon rainforest. It presents a unique environmental mosaic, where high biodiversity in natural vegetation coexists with intense agricultural activity, leading to approximately 25% of the territory being occupied by anthropogenic land use [14]. Despite this, the region remains one of the least densely populated areas within the *Cerrado*, preserving a substantial portion of native vegetation.

The *Bico do Papagaio* was purposefully selected as a representative pilot region for the initial integration of MSI and SAR data in LULC classification studies. This ecoregion encapsulates many of the biome's broader challenges, particularly the intricate

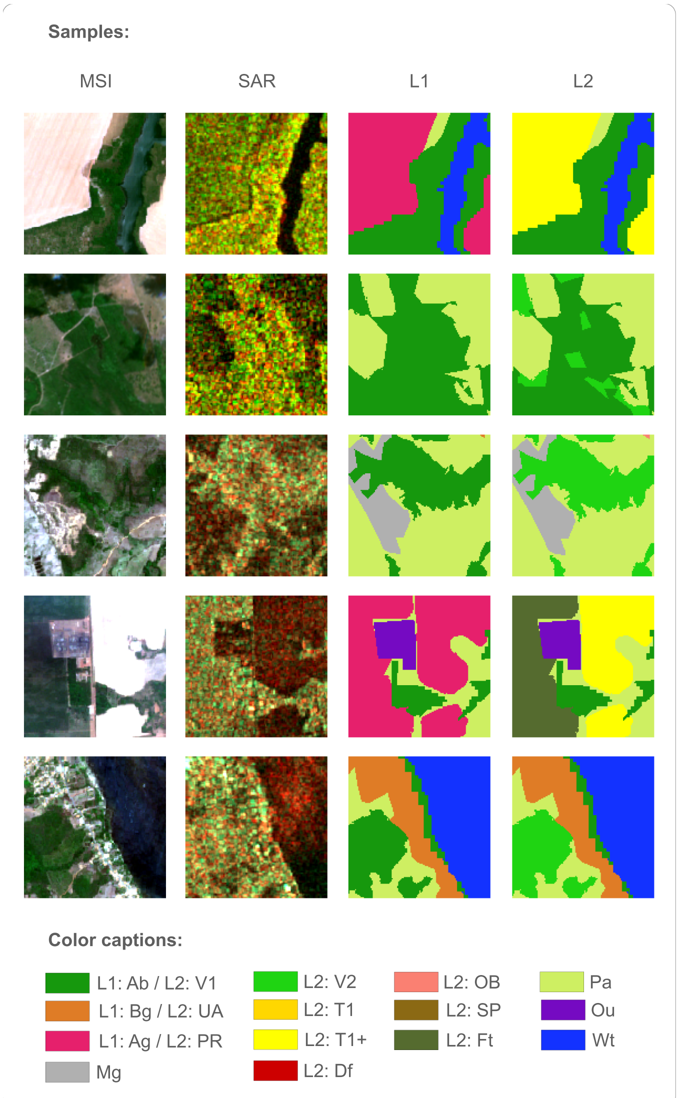

Fig. 2. Samples of the Bico do Papagaio ecorregion are displayed alongside their respective semantic at both hierarchical levels L1 and L2, which classes are represented in different colours.

heterogeneity between natural formations and human-modified landscapes, which are common across various parts of the *Cerrado*, as illustrated in Fig. 2. Therefore, although geographically limited, the dataset captures relevant and transferable classification complexities. This regional scope allows the development and assessment of robust multimodal DL methods in a real world context, laying the foundation for future expansions to other regions of the biome.

A. Data Source

CerraData-4 MM integrates SAR and multispectral rasters from the S1 and S2 satellites, respectively. The images are collected from the Copernicus platform[1] relating to year 2022. Days with minimal cloud cover are prioritized to ensure optimal visibility of the targets. For S1, seven scenes are used, each featuring dual-polarization (VV and VH) bands from the Level-1 ground range detected product. These bands have a spatial resolution of 10 meters. For S2, 18 scenes are used at Level-2 A. Each scene includes four bands with 10 meters spatial resolution (B2, B3, B4, and B8), six bands with a 20 meters spatial resolution (B5, B6, B7, B8a, B11, and B12), and two atmospheric bands with a 60 m spatial resolution (B9 and B10).

B. Reference Data

The reference data are annotated founded on TerraClass *Cerrado* 2022 map,[2] which delineates a comprehensive range of LULC categories, characterized by vegetation structure, land management practices, and anthropogenic influences. TerraClass map comprises 15 classes, including primary and secondary natural vegetation, five types of agricultural land, water bodies, and deforested areas. It is a highly accurate product, meticulously developed by experts in LULC classification. The map is built up on Sentinel-2 scenes, which provides a spatial resolution of 10 m. The classification process involves manual annotation for urban-related classes [urban areas, built structures, mining, and other uses (OU)], and time series analysis using DL models for agricultural classes, e.g., temporary agriculture with a single cycle, multiple cycles, perennial and semi-perennial agriculture. For secondary natural vegetation and pastures, the self-organizing maps technique is used to aid in the annotation phase.

C. Data Preprocessing

The source data undergoes three processing steps in the SNAP tool.[3] First, all bands are resampled to a resolution of 10 m using bilinear interpolation. Second, the bands are stacked to generate rasters comprising 12 channels for optical images and two channels for SAR images. Third, the scenes are reprojected to the EPSG 4326 coordinate system using bilinear interpolation and cropped into patches with dimensions of 128×128 pixels. Patches containing NaN or Null values are subsequently discarded.

Reference data is structured into two hierarchical levels of classes, providing macro and micro representations of land use/land cover (LULC). Level 1 (L1) comprises seven classes, each belonging to a class group that offers a macro view of subclasses within Level 2 (L2). For instance, the class "Agriculture" (Ag) encompasses five classes represented at Level 2: "Perennial Agriculture" (Pr), "Semi-perennial Agriculture" (SP), "Temporary Agriculture of 1 cycle" (T1), "Temporary Agriculture of 1+ cycle" (T1+), and "Silviculture" (Ft). Table II elucidates the definitions, structures, and relationships of each class at both levels. The hierarchical structure of classes at both levels is governed by the TerraClass map and defined by their semantic connections. For example, the class "Arboreal" (Ab) at Level 1 has a direct relationship with the classes "Primary Natural vegetation" (V1) and "Secondary Natural vegetation" (V2), as both classes share the same definition but differ in terms

[1]Copernicus Browser: https://dataspace.copernicus.eu/

[2]TerraClass 2022 map: https://www.terraclass.gov.br/download-de-dados

[3]SNAP tool: https://step.esa.int/main/toolboxes/snap/

TABLE II
HIERARCHICAL RELATIONSHIP OF THE CLASSES, ALONG WITH THE PERCENTAGE OF SAMPLES PER CLASS AND THEIR DESCRIPTIONS BASED ON TERRACLASS DEFINITIONS

Level 1	**Description**	**Number of Samples (%)**	**Level 2**	**Description**	**Number of Samples (%)**
Pasture (Pa)	Is composed of herbaceous forage species, either cultivated or naturalized, and may include varying degrees of shrub or tree cover	35.25	Pasture (Pa)	Is composed of herbaceous forage species, either cultivated or naturalized, and may include varying degrees of shrub or tree cover	35.25
Arboreal (Ab)	Comprises both primary and secondary natural vegetation	58.43	Primary Natural Vegetation (V1)	Refers to native formations, ranging from grasslands to dense savanna woodlands (cerradão), that have not experienced total vegetation removal since the beginning of monitoring	52.21
			Secondary Natural Vegetation (V2)	Represents regenerated areas, e.g., primarily savanna forest, which have undergone complete deforestation at some point in the historical record	6.22
Agriculture (Ag)	Encompasses crop areas, such as those of perennial to temporary agriculture	4.43	Perennial (Pr)	Comprises long-cycle crops such as coffee	0.07
			Semi-perennial (SP)	Consists of crops like sugarcane with production cycles extending beyond one year	0.15
			Temporary agriculture of 1 cycle (T1)	Refers to short-cycle crops, predominantly grains and fibers, cultivated once per year	0.27
			Temporary agriculture of 1+ cycle (T1+)	Involves multiple harvests within the same year	1.15
			Forestry (Ft)	(Silviculture) Includes commercial plantations of native or exotic tree species, typically forming homogeneous arboreal stands	1.49
			Deforestation (Df)	Identifies regions where natural vegetation was entirely cleared during 2022	1.30
Mining (Mg)	Characterized by exposed soils and substantial landscape alteration due to mineral extraction	0.01	Mining (Mg)	Characterized by exposed soils and substantial landscape alteration due to mineral extraction.	0.01
Building (Bg)	Refers to areas with high and low densities of buildings, such as cities, farms, and villages.	0.55	Urban area (UA)	Comprehend cities, towns, and other settlements marked by infrastructure development and building density	0.45
			Other Built area (OB)	Comprises areas with low population density, such as farms, agro-industrial complexes, villages, hydroelectric plants, dams, and other infrastructure.	0.10
Water body (Wt)	Encompass both natural and artificial aquatic features, including reservoirs, ponds, and aquaculture facilities	1.32	Water body (Wt)	Encompass both natural and artificial aquatic features, including reservoirs, ponds, and aquaculture facilities	1.32
Other Uses (OU)	Encompasses other land use features not included in this dataset, such as fire scars, rock outcrops, beaches, clouds, and cloud shadows	0.01	Other Uses (OU)	Encompasses other land use features not included in this dataset, such as fire scars, rock outcrops, beaches, clouds, and cloud shadows	0.01

of vegetation regeneration degree. Table II provides further details on this matter.

D. CerraData-4 MM

CerraData-4 MM consists of paired SAR and MSI images along with semantic maps for the levels L1 and L2, presented in Fig. 2. Each hierarchical level contains 30 322 image patches with a spatial resolution of 10 m. This dataset encompasses a diverse set of LULC classes representative of the rich biodiversity of *Bico do Papagaio* region, *Cerrado* biome. The semantic map provides a macro (L1) and micro (L2) perspectives on the region's vegetation structure and land cover types, as detailed in Table II. The dataset presents two primary challenges for classification models, e.g., DL models: 1) similarity among certain categories, such as between primary and secondary natural vegetation; and 2) class imbalance, with a disproportionate number of samples across different categories, e.g., OU class.

IV. Evaluation

To assess the learnability of the dataset, two baseline DL architectures are considered: U-Net [19] and TransNuSeg [24]. These models are chosen to reflect the dominant paradigms in semantic segmentation: convolutional architectures and Transformer-based designs. U-Net, as a canonical CNN model, serves as a strong STL baseline due to its proven effectiveness in dense prediction tasks, including remote sensing. Conversely, TransNuSeg represents recent advances in ViT-based MTL architectures, capturing long-range dependencies and offering a modern benchmark for evaluating the dataset's capacity to support complex, multioutput learning. By incorporating these diverse models, the study provides a comprehensive evaluation of the dataset's potential, spanning both convolutional and Transformer-based methodologies. The training and testing phase are carried out in a workstation equipped with an NVIDIA GeForce RTX 3060 GPU, comprising 12 GB VRAM, CUDA 12.2, 62 GB of RAM, and a 13th Gen Intel Core i7-13700 processor, 24 threads, 16 cores. This hardware configuration ensures sufficient computational capacity to support the training of both convolutional and Transformer-based DL models employed in this evaluation setup.

Based on the CerraData-4 MM reference data, modifications are made to the TransNuSeg architecture to support two specific tasks: semantic segmentation and edge segmentation. In contrast, the original architecture of U-Net is trained without weight initialization. Given the dataset's multimodal nature, an ablation study is conducted in which all models are trained under three distinct scenarios for both L1 and L2 levels: 1) using only MSI, 2) using only SAR data, and 3) using a concatenated combination of both modalities. No preprocessing is applied to the input images. A weighted loss function is employed to address class imbalance, with weights derived from the sample distribution of the reference masks (see Table II). Models performance are evaluated using the metrics: the macro-averaged F1-score, mean Intersection over Union (mIoU), and confusion matrix.

The configuration of hyperparameters employed in the experiments encompassed both shared settings across all models and architecture-specific adjustments. For all models, training is conducted using the Adam optimizer with a learning rate of 0.001, a batch size of 32, and a total of 100 training epochs. Regarding individual architectures, the TransNuSeg model utilized a composite loss function combining Dice and cross-entropy (CE) losses, with a Swin-Transformer as its encoder backbone and a parameter sharing ratio of 0.8. The U-Net architecture employes the CE loss function and used a ResNet-50 as the encoder. These tailored hyperparameter settings ensured a fair comparison while leveraging each model's architectural strengths in the context of multimodal LULC classification.

A. Baseline Results

Fig. 3 summarizes the overall performance of the baseline models across various training configurations, reporting the F1-score and the mIoU for each scenario. The results consider both weighted (W) and nonweighted (NW) loss functions applied to different modality inputs, i.e., MSI, SAR, and MSI + SAR, and hierarchical levels, i.e., L1 and L2.

TransNuSeg consistently outperforms U-Net in nearly all scenarios at the L1 level. Its ViT-based architecture achieves higher scores in both F1 and mIoU metrics, particularly in settings involving the fusion of MSI and SAR data. For instance, in the Concat-NW-L1 scenario, TransNuSeg reaches an F1-score of 57.60%, outperforming its MSI-NW-L1 and SAR-NW-L1 variants, 57.45% and 38.94%, respectively. U-Net exhibits lower and more variable performance across all configurations. Training exclusively on MSI data yields significantly better results than training with SAR only. This pattern is observed for both models and is attributed to the richer spectral representation in MSI. Combining MSI and SAR modalities leads to the highest scores in most cases, reinforcing the advantage of multimodal input.

Performance declines notably when models are evaluated at the second hierarchical level (L2). TransNuSeg's F1-score drops from 57.45% to 53.17% and mIoU from 48.94% to 41.90% in the MSI-NW scenario. U-Net experiences a more drastic drop in Concat-NW-L2, with an F1-score of 18.16% and mIoU of 15.84%. These reductions are consistent with the increase in the complexity of the classification in L2, which involves more classes and a higher imbalance. The application of class-weighted loss has mixed effects. For TransNuSeg, it often results in slight performance degradation, as seen in MSI-W-L1 compared to MSI-NW-L1, indicating that weighting may impact the model's treatment of majority classes. U-Net, in contrast, shows modest improvements in specific settings, such as SAR-W-L1.

Figs. 4 and 5 display the confusion matrices for models trained on weighted MSI+SAR data at L1 and L2, respectively. These matrices illustrate the distribution of class-wise errors and reveal persistent misclassifications among visually similar or underrepresented classes. Qualitative predictions in Figs. 6 and 7 provide visual confirmation of the numerical trends. TransNuSeg outputs tend to be more spatially coherent and refined, especially in L1. Weighted loss training enhances the detection of rare classes, such as "OU" and "Mg," although it introduces more errors in majority-class regions. U-Net shows

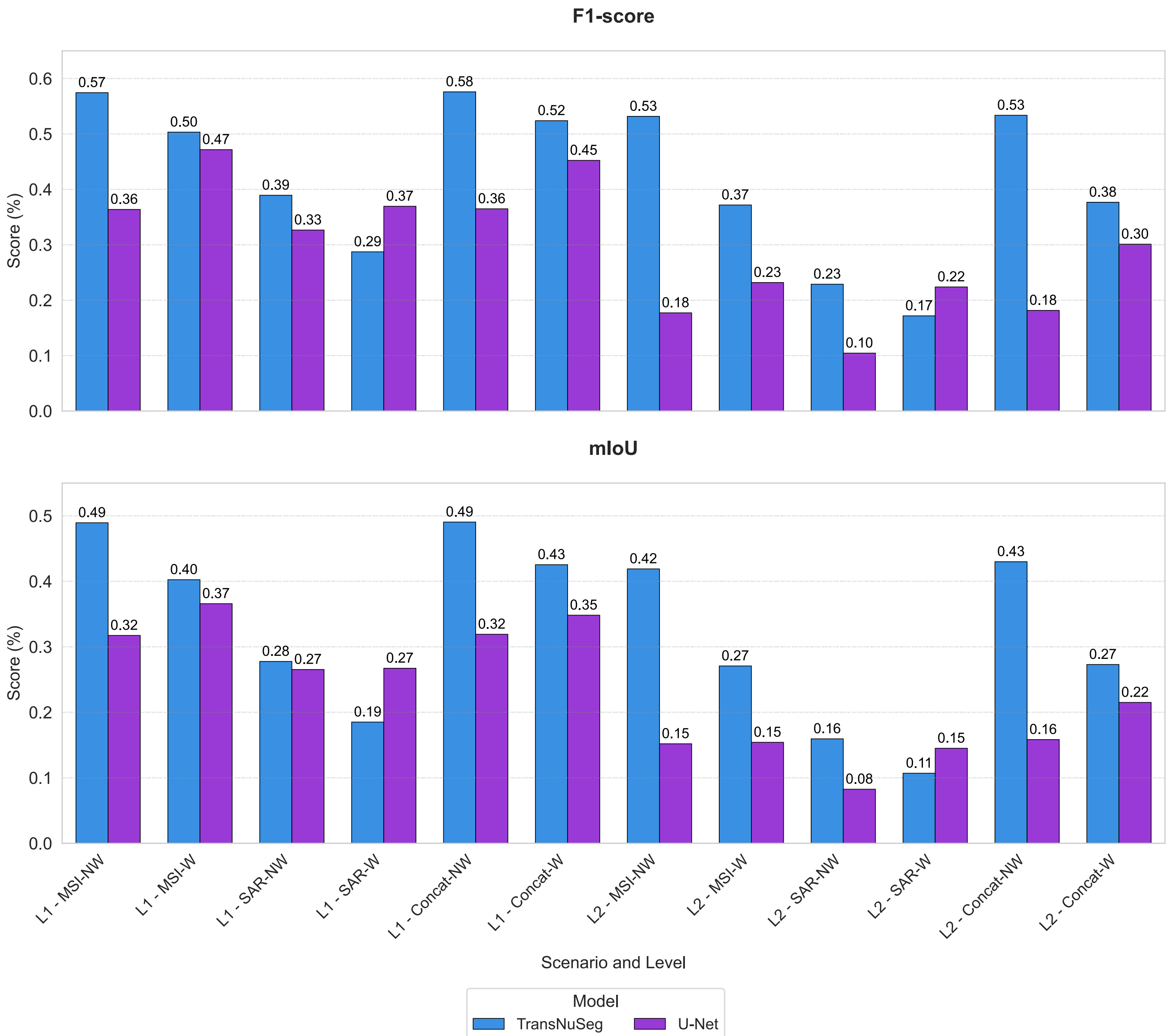

Fig. 3. Performance metrics (F1-score and IoU) of TransNuSeg and U-Net models across different input modalities (MSI, SAR, and MSI+SAR) for both hierarchical levels (L1 and L2).

increased confusion between similar classes like “Pa” and “Ag,” particularly in unweighted scenarios.

B. Discussion

The experimental results presented in the previous section offer critical insights into the capabilities and limitations of the evaluated segmentation models, particularly when dealing with hierarchical labelled multimodal datasets like CerraData-4 MM. These models’ performance provides an overview of the challenges involved in classifying LULC in the Cerrado biome, highlighting the need for more efficient approaches to manage the imbalance and diversity of vegetation types. In light of this, CerraData-4 MM plays a key role in revealing these challenges and in promoting the development and adoption of more effective methods to address them.

The outstanding performance of the TransNuSeg model over U-Net in most scenarios, especially at the first hierarchical level (L1), highlights the advantages of transformer-based architectures in learning contextual and spatial relationships. This advantage is especially pronounced when dealing with complex and heterogeneous classes, as the attention mechanisms intrinsic to vision transformers allow for better generalization and spatial coherence. Conversely, U-Net, despite being a robust convolutional architecture, tends to struggle with interclass ambiguities and produces more fragmented segmentations, particularly for minority classes and in higher complexity scenarios, such as L2.

Consistent improvement observed when combining MSI and SAR modalities reinforces the importance of multimodal data fusion in remote sensing tasks. While MSI data provides rich spectral information, SAR data offers complementary structural

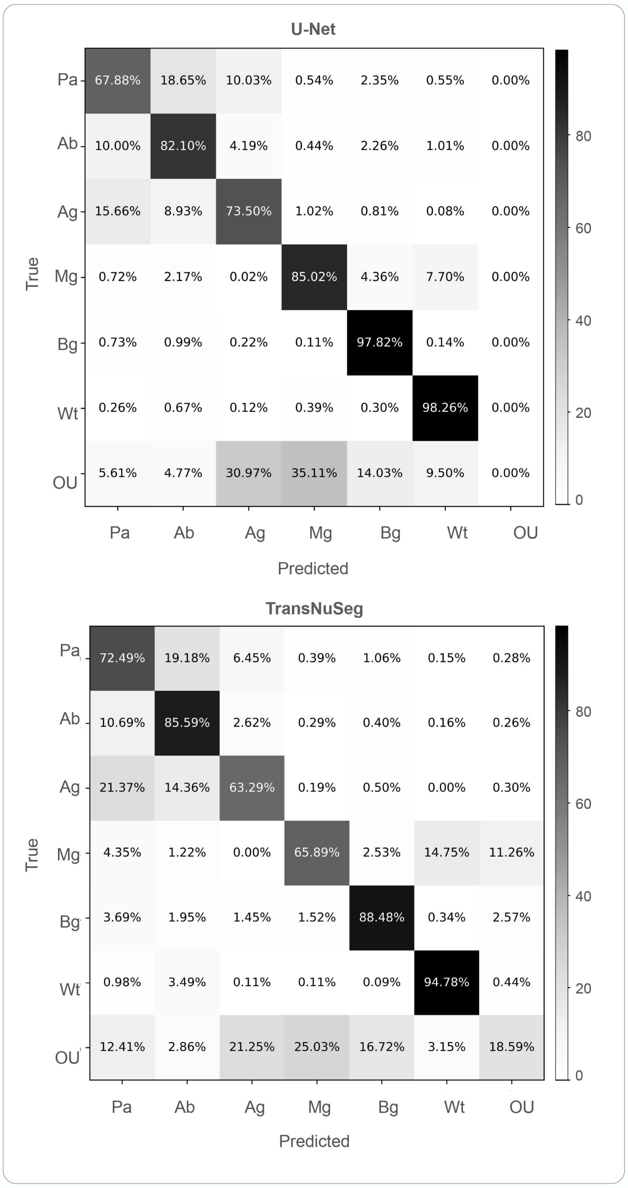

Fig. 4. Confusion matrix for weighted MSI+SAR L1 data-trained models.

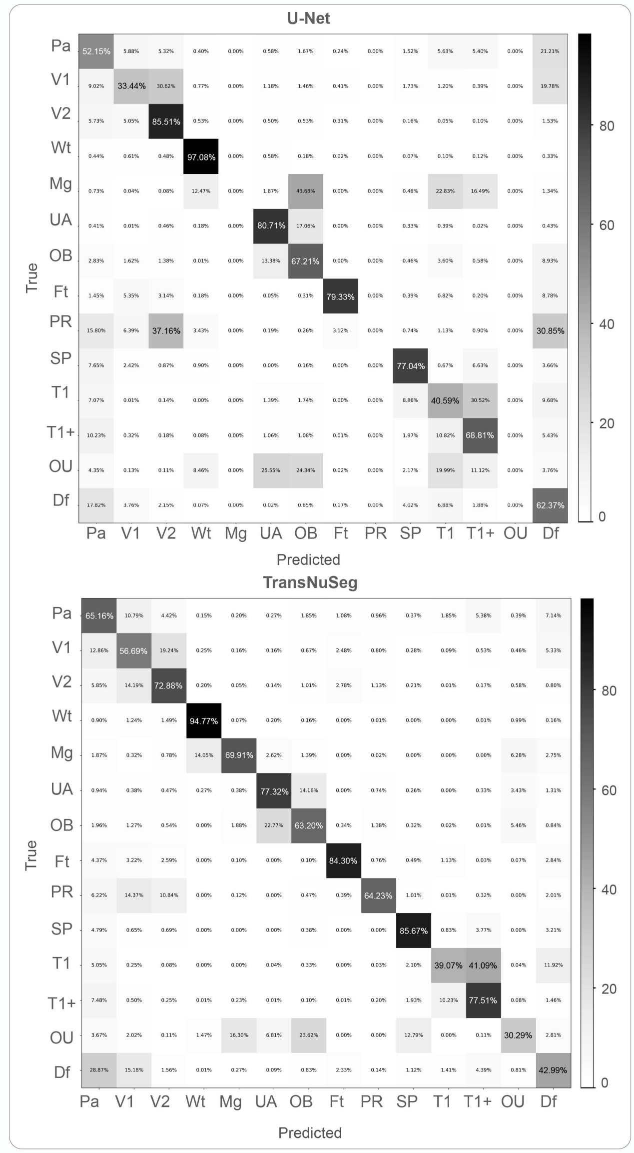

Fig. 5. Confusion matrix for weighted MSI+SAR L2 data-trained models.

features. However, SAR-only scenarios result in limited performance, underscoring the need for careful modality selection and combination depending on the target classes. The concatenation of both modalities leads to significant performance gains, confirming the potential of joint representations to improve model robustness.

The application of class-weighted loss functions introduces a nuanced tradeoff. Although overall metric scores, such as F1 and mIoU may slightly decline, the use of weights proves effective in enhancing the model's sensitivity to underrepresented classes. This is particularly valuable in real-world applications, where class imbalance is a persistent challenge. TransNuSeg, benefit less from weighted loss in aggregate metrics, shows improved detection of minority classes without severely compromising the segmentation quality of majority ones. U-Net, on the other hand, benefits more variably, with some improvement in specific cases, such as SAR-based training, however still exhibits limitations in representing rare categories consistently. The confusion matrices and qualitative prediction samples further illustrate the inherent difficulty in classifying fine-grained L2 categories. As the number of classes increases and their distributions become more skewed, model performance degrades. This degradation is expected due to higher interclass similarity and limited sample availability for rare classes. Nevertheless, both models, especially TransNuSeg, demonstrate a promising capacity to identify minority classes under weighted training, albeit with some misclassification in visually similar regions.

Based on qualitative analysis, important contrasts between the models are observed. TransNuSeg generally produces smoother, more cohesive segmentations with sharper boundaries, especially in homogeneous regions. In contrast, U-Net's outcomes often show over-segmentation and spatial inconsistency, particularly in complex scenes. This reinforces the notion that

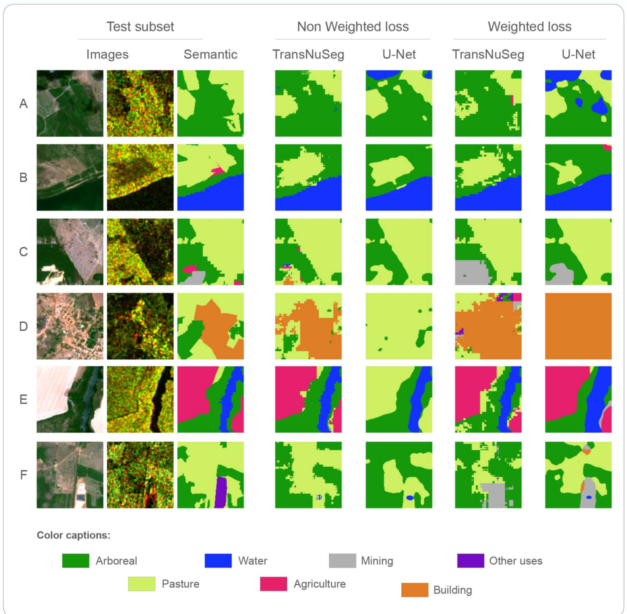

Fig. 6. Samples of predictions by U-Net and TransNuSeg trained with both modalities in L1.

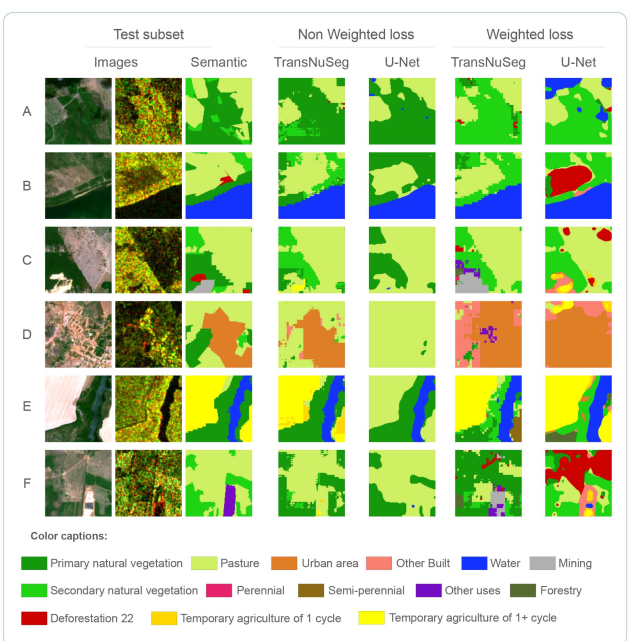

Fig. 7. Samples of predictions by U-Net and TransNuSeg trained with both modalities in L2.

attention-based models are more adept at capturing global dependencies and subtle visual cues essential for distinguishing between closely related classes. Therefore, the results support the assumption that transformer-based architectures, when combined with multimodal data and appropriate training strategies, such as class-weighted losses, provide a promising direction for improving semantic segmentation performance in imbalanced and hierarchically structured datasets. However, challenges remain, particularly in enhancing the detection of minority classes and maintaining performance consistency across hierarchical levels. Future work could explore more sophisticated fusion strategies, adaptive weighting schemes, and data augmentation techniques tailored to underrepresented classes.

V. Final Remarks

This article introduces CerraData-4 MM, a multimodal dataset developed for LULC classification in the *Cerrado* biome. The dataset integrates SAR and MSI imagery from Sentinel-1 and Sentinel-2, with patches at a spatial resolution of 10 meters. Provides two hierarchical semantic levels of classification, comprising seven and 14 categories, along with corresponding edge information. CerraData-4 MM encompasses the *Bico do Papagaio* ecoregion, which poses unique classification challenges due to its mosaic of natural vegetation, anthropogenic activities, and agricultural subcategories. This dataset presents a real-world challenge due to the prevalence of class imbalance, which may require preprocessing techniques and models capable of learning from heterogeneous and visually ambiguous data.

The experimental results underscore these challenges, while also highlighting the opportunities for innovation in DL models. Approaches, such as U-Net and TransNuSeg, demonstrate strong performance in majority classes, however struggle with minority categories and subtle spectral differences within agricultural subclasses. The weighted loss strategy improves the representation of underrepresented classes. However, consistent misclassifications in visually similar classes emphasize the necessity for advanced techniques, such as attention mechanisms, multimodal fusion, and hierarchical learning frameworks.

For the next version of CerraData, it is expected to include other Cerrado ecoregions in order to further increase the diversity of samples, with regard to visual patterns. Given the diversity of vegetation domains, it is planned to include more natural classes, e.g., rupestrian savannas and gallery forests, thus encouraging the mapping and historical reporting of environmental preservation of such groups of vegetation formation. The integration of other modalities, e.g., DEM, for height estimation tasks can be considered to promote support for multitask learning models.

Acknowledgment

This research was conducted within the *Laboratório de Inteligência ARtificial para Aplicações AeroEspaciais e Ambientais* (LIAREA) related to the project *Classificação de imagens e dados via redes neurais profundas para múltiplos domínios* (*IDeepS*). The *IDeepS* project repository is publicly available online.[4] The authors would like to thank Brazilian Federal Agency for Support and Evaluation of Graduate Education for their support of a research exchange at the German Aerospace Center (DLR) in Oberpfaffenhofen. The authors would also like to thank the support of the *Laboratório Nacional de Computação Científica* (LNCC, MCTI/Brazil) for providing access to the computational resources of the SDumont supercomputer.

[4][Online]. Available: https://github.com/vsantjr/IDeepS

REFERENCES

[1] M. Schmitt, L. Hughes, C. Qiu, and X. Zhu, "SEN12 MS - a curated dataset of georeferenced multi-spectral sentinel-1/2 imagery for deep learning and data fusion," *ISPRS Ann. Photogrammetry, Remote Sens. Spatial Inf. Sci.*, vol. 4, pp. 153–160, 2019.

[2] K. Nogueira et al., "Towards vegetation species discrimination by using data-driven descriptors," in *Proc. 9th IAPR Workshop Pattern Recogniton Remote Sens.* (PRRS), 2016, pp. 1–6.

[3] R. Wenger, A. Puissant, J. Weber, L. Idoumghar, and G. Forestier, "Multisenge: A multimodal and multitemporal benchmark dataset for land use/land cover remote sensing applications," *ISPRS Ann. Photogrammetry, Remote Sens. Spatial Inf. Sci.*, vol. 3, pp. 635–640, 2022.

[4] J. Hu et al., "MDAS: A new multimodal benchmark dataset for remote sensing," *Earth Syst. Sci. Data*, vol. 15, no. 1, pp. 113–131, 2023. [Online]. Available: https://essd.copernicus.org/articles/15/113/2023/

[5] C. Robinson et al., "Global land-cover mapping with weak supervision: Outcome of the 2020 IEEE GRSS data fusion contest," *IEEE J. Sel. Topics Appl. Earth Observ. Remote Sens.*, vol. 14, pp. 3185–3199, 2021.

[6] G. Baier, A. Deschemps, M. Schmitt, and N. Yokoya, "Synthesizing optical and SAR imagery from land cover maps and auxiliary raster data," *IEEE Trans. Geosci. Remote Sens.*, vol. 60, 2022, Art no. 4701312.

[7] V. S. f. Garnot, L. Landrieu, and N. Chehata, "Multi-modal temporal attention models for crop mapping from satellite time series," *ISPRS J. Photogrammetry Remote Sens.*, vol. 187, pp. 294–305, 2022.

[8] R. Hänsch et al., "Data fusion contest 2022 (dfc2022)," *IEEE Dataport*, 4, January 2022, doi: 10.21227/rjv6-f516

[9] A. Garioud, S. Peillet, E. Bookjans, S. Giordano, and B. Wattrelos, "Flair-1: Semantic segmentation and domain adaptation dataset," 2022. [Online]. Available: https://rgdoi.net/10.13140/RG.2.2.30183.73128/1

[10] Y. Li, Y. Zhou, Y. Zhang, L. Zhong, J. Wang, and J. Chen, "Dkdfn: Domain knowledge-guided deep collaborative fusion network for multimodal unitemporal remote sensing land cover classification," *ISPRS J. Photogrammetry Remote Sens.*, vol. 186, pp. 170–189, 2022.

[11] D. Hong et al., "Cross-city matters: A multimodal remote sensing benchmark dataset for cross-city semantic segmentation using high-resolution domain adaptation networks," *Remote Sens. Environ.*, vol. 299, 2023, Art. no. 113856. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0034425723004078

[12] B. B. N. Strassburg et al., "Moment of truth for the Cerrado hotspot," *Nature Ecol. Evol.*, vol. 1, 1, 3, 2017.

[13] E. L. BOLFE, E. E. SANO, and S. K. CAMPOS, "Dinâmica agrícola no cerrado: Análises e projeções," *Emprapa*, vol. 1, 2020. [Online]. Available: https://ainfo.cnptia.embrapa.br/digital/bitstream/item/212381/1/LV-DINAMICA-AGRICOLA-CERRADO-2020.pdf

[14] J. C. de Oliveira Santana and M. F. Simon, "Plant diversity conservation in an agricultural frontier in the Brazilian Cerrado," *Biodiversity Conservation*, vol. 31, 2021.

[15] R. S. OLIVEIRA et al., "Deep root function in soil water dynamics in cerrado savannas of central Brazil," *Funct. Ecol.*, vol. 19, no. 4, pp. 574–581, 2024. [Online]. Available: https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/j.1365-2435.2005.01003.x

[16] "Terraclass mostra a cobertura e o uso da terra no bioma cerrado," *Empresa Brasileira De Pesquisa Agropecuária (Embrapa)*, 2022. [Online]. Available: https://www.embrapa.br/busca-de-noticias/-/noticia/77150778/terraclass-mostra-a-cobertura-e-o-uso-da-terra-no-bioma-cerrado

[17] C. A. de Almeida et al., "High spatial resolution land use and land cover mapping of the Brazilian legal Amazon in 2008 using Landsat-5/tm and modis data," *Acta Amazonica*, vol. 46, pp. 291–302, 2016.

[18] J. Yao, B. Zhang, C. Li, D. Hong, and J. Chanussot, "Extended vision transformer (ExViT) for land use and land cover classification: A multimodal deep learning framework," *IEEE Trans. Geosci. Remote Sens.*, vol. 61, 2023, Art. no. 5514415.

[19] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional networks for biomedical image segmentation," in *Proc. Med. Image Comput. Comput.-Assisted Intervention–MICCAI* 2015, 2015, pp. 234–241.

[20] A. K. Neves, T. S. Körting, L. M. G. Fonseca, A. R. Soares, C. D. Girolamo-Neto, and C. Heipke, "Hierarchical mapping of Brazilian Savanna (Cerrado) physiognomies based on deep learning," *J. Appl. Remote Sens.*, vol. 15, 2021, Art. no. 0 44504.

[21] L. M. Fonseca et al., "Pattern recognition and remote sensing techniques applied to land use and land cover mapping in the Brazilian Savannah," *Pattern Recognit. Lett.*, vol. 148, pp. 54–60, 2021.

[22] M. de Souza Miranda, V. A. de Santiago Júnior, T. S. Körting, E. C. dos S. Monteiro, and J. Q. da Silva, "Ai4luc: Deep learning and automated mask labelling to support land use and land cover mapping in the cerrado biome," *Remote Sens. Lett.*, vol. 15, no. 8, pp. 850–860, 2024.

[23] M. D. S. Miranda, L. F. A. e. Silva, S. F. D. Santos, V. A. de Santiago Júnior, T. S. Körting, and J. Almeida, "A high-spatial resolution dataset and few-shot deep learning benchmark for image classification," in *Proc. 35th SIBGRAPI Conf. Graph., Patterns Images (SIBGRAPI)*, 2022, pp. 19–24.

[24] Z. He, M. Unberath, J. Ke, and Y. Shen, "Transnuseg: A lightweight multi-task transformer for nuclei segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.- Assist. Intervention*, 2023, pp. 206–215.

Mateus de Souza Miranda received the master's degree in applied computing from the the National Institute for Space Research, São José dos Campos, Brazil, where he is currently working toward the Ph.D. degree in applied computing.

His research interests include multitask learning and self-supervised deep learning models applied for land use and land cover classification through satellite images.

Ronny Hänsch received the undergraduation, M.Sc., and Ph.D. degrees from TU Berlin, Berlin, Germany, in 2005, 2007, and 2014, respectively.

2002–2005 Undergraduate degree TU Berlin, 2005–2007 Graduate degree (MSc) TU Berlin, 2007–2014 Graduate degree (PhD) TU Berlin, 05.2019–present: Postdoctoral research fellow, Department SAR Technology German Aerospace Center (DLR), Oberpfaffenhofen, Germany.

Since 2019, he has been a Postdoctoral Research Fellow with the Department SAR Technology, German Aerospace Center (DLR), Oberpfaffenhofen, Germany. He also serves as the lead of the "Machine Learning" team with the Department of SAR Technology, German Aerospace Center, Cologne, Germany. His research interests include advanced machine learning techniques, including deep learning and random forests, applied to remote sensing with a particular emphasis on polarimetric synthetic aperture radar imagery, and innovative domains, such as self-supervised learning and cross-modal learning.

Valdivino Alexandre de Santiago Júnior received the master's degree in electrical engineering from the *Universidade Federal do Ceará*, Fortaleza, Brazil, in 1999, and the Ph.D. degree from the postgraduate program in applied computing of the *Instituto Nacional de Pesquisas Espaciais* (INPE), São José dos Campos, Brazil, in 2011.

He also holds the position of Senior Technologist and is the leading Researcher with the *Laboratório de Inteligência ARtificial para Aplicações AeroEspaciais e Ambientais*, INPE. His research interests include artificial intelligence, machine learning, deep learning, large language models, remote sensing, environmental applications, systems/software optimization, and aerospace systems (satellites, drones).

Thales Sehn Körting received the B.S. degree in computer engineering from the Federal University of Rio Grande (FURG), Rio Grande, Brazil, in 2005, and the M.S. degree in applied computing and Ph.D. degree in remote sensing from National Institute for Space Research, São José dos Campos, Brazil, in 2007 and 2012, respectively.

His research interests include remote sensing image segmentation, multitemporal analysis, image classification, data mining algorithms, and artificial intelligence.

Erison Carlos dos Santos Monteiro received the bachelor's degree in biological sciences and the master's degree in ecology and conservation from the Federal University of Mato Grosso do Sul, Campo Grande, Brazil, in 2008 and 2013, respectively, and the Ph.D. degree in ecology and biodiversity from São Paulo State University Júlio de Mesquita Filho, São Paulo, Brazil, in 2020.

He was a Visiting Researcher with the Senckenberg Biodiversity and Climate Research Centre, Frankfurt, Germany, and investigated how human activities on landscapes affect species interactions and how the extinction of these interactions can influence the evolutionary aspects of biodiversity. Recently, he was appointed to a Technician position with the Brazilian National Institute for Space Research, São José dos Campos, Brazil, where he coordinate the INPE team for TerraClass Cerrado—a project mapping the land cover of the Cerrado biome in Brazil.

Dr. Monteiro was the recipient of the fellowship from Senckenberg Gesellschaft für Naturforschung in 2020.

Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) - ROR identifier: 00x0ma614